Experimental methodology for data science

OR: Stressing less about our (questionable) experimental choices

---

Carlos Ramisch
first.last@lis-lab.fr

M2 IAAA - based on the course *Zen Research*
By Carlos Ramisch and Manon Scholivet

Who are we?

Manon Scholivet

- PhD defended in 2021
- An enthusiast for neat and tidy science
- Scientific methodology learnt by trial and error

Who are we?

Manon Scholivet

- PhD defended in 2021
- An enthusiast for neat and tidy science
- Scientific methodology learnt by trial and error

Carlos Ramisch

- PhD defended in 2012
- A big fan of idiomatic expressions
- Scientific methodology learnt by trial and error

Course goals

Goal

Share our know-how in experimental methodology for data science

$\rightarrow$ Key notions, recommended practices, avoidable traps...

Collaboratively build an ideal of scientific methodology

Course goals

Goal

Share our know-how in experimental methodology for data science

→ Key notions, recommended practices, avoidable traps...

Collaboratively build an ideal of scientific methodology

- Experiments are a central component of data science
- Methodology and experimental design are often neglected

  → Shaky conclusions, stressful writing, negative feedback...
- Special time: study notions often used without understanding

  → Greatly influence our work's quality

Course history

- ED 184 version (2022 and 2023):

  `https://pageperso.lis-lab.fr/carlos.ramisch/?page=recherchezen`

- ESSLLI version (2024):

  `https://gitlab.com/zenresearch/esslli2024/`

- M2 IAAA version (2024-present):

  $\rightarrow$ Focus on more applied data science

  $\rightarrow$ Course materials on Ametice

Wooclap and GDocs: please, send an email if you want to modify/delete your data

Course outline - M2 IAAA version

Lesson 1: Defining a research question

Lesson 2: Reading a scientific paper (evaluation!)

Lesson 3: Preparing the data and experiments

Lesson 4: Evaluating the models

Lesson 5: Analysing our results (evaluation!)

Lesson 6: Sharing our findings

Methodology
Lesson 1: Defining a research question

Carlos Ramisch
first.last@lis-lab.fr

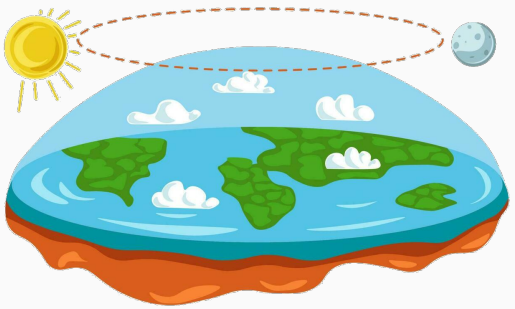
M2 IAAA - based on the course *Zen Research*
By Carlos Ramisch and Manon Scholivet

Outline

Wooclap

Wooclap time!

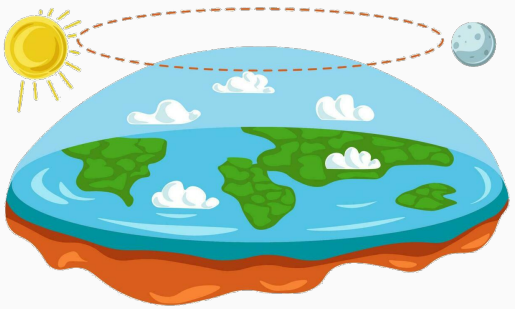
What is science?



- Scientific statements are those that prove to be certain

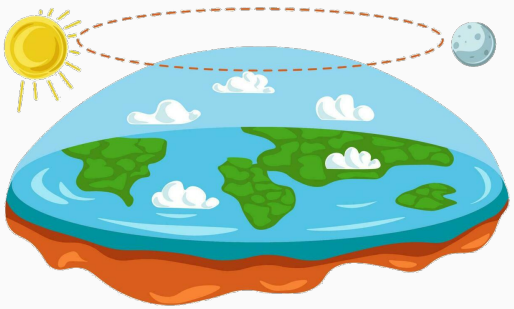
  $\rightarrow$ We were certain that the sun revolved around the Earth, and yet...

What is science?



- Scientific statements are those that prove to be certain
  - $\rightarrow$ We were certain that the sun revolved around the Earth, and yet...
- Science = scientific method
  - $\rightarrow$ Scientific method also used in journalism, flat Earth conspiracy...
  - $\rightarrow$ Methods evolve, e.g. use of control groups in clinical trials

What is science?



- Scientific statements are those that prove to be certain
  - → We were certain that the sun revolved around the Earth, and yet...
- Science = scientific method
  - → Scientific method also used in journalism, flat Earth conspiracy...
  - → Methods evolve, e.g. use of control groups in clinical trials
- Formulate testable hypotheses
  - → So astrology is science, but string theory is not!

Wooclap

Wooclap time!

Data science

Is **data science** a science?

- Data scientists develop and evaluate machine learning models
- How do we evaluate these models?

Data science

Is **data science** a science?

- Data scientists develop and evaluate machine learning models
- How do we evaluate these models?
- Carry out experiments!

Data science experiments

Regardless of being a "real" science, data science is scientific

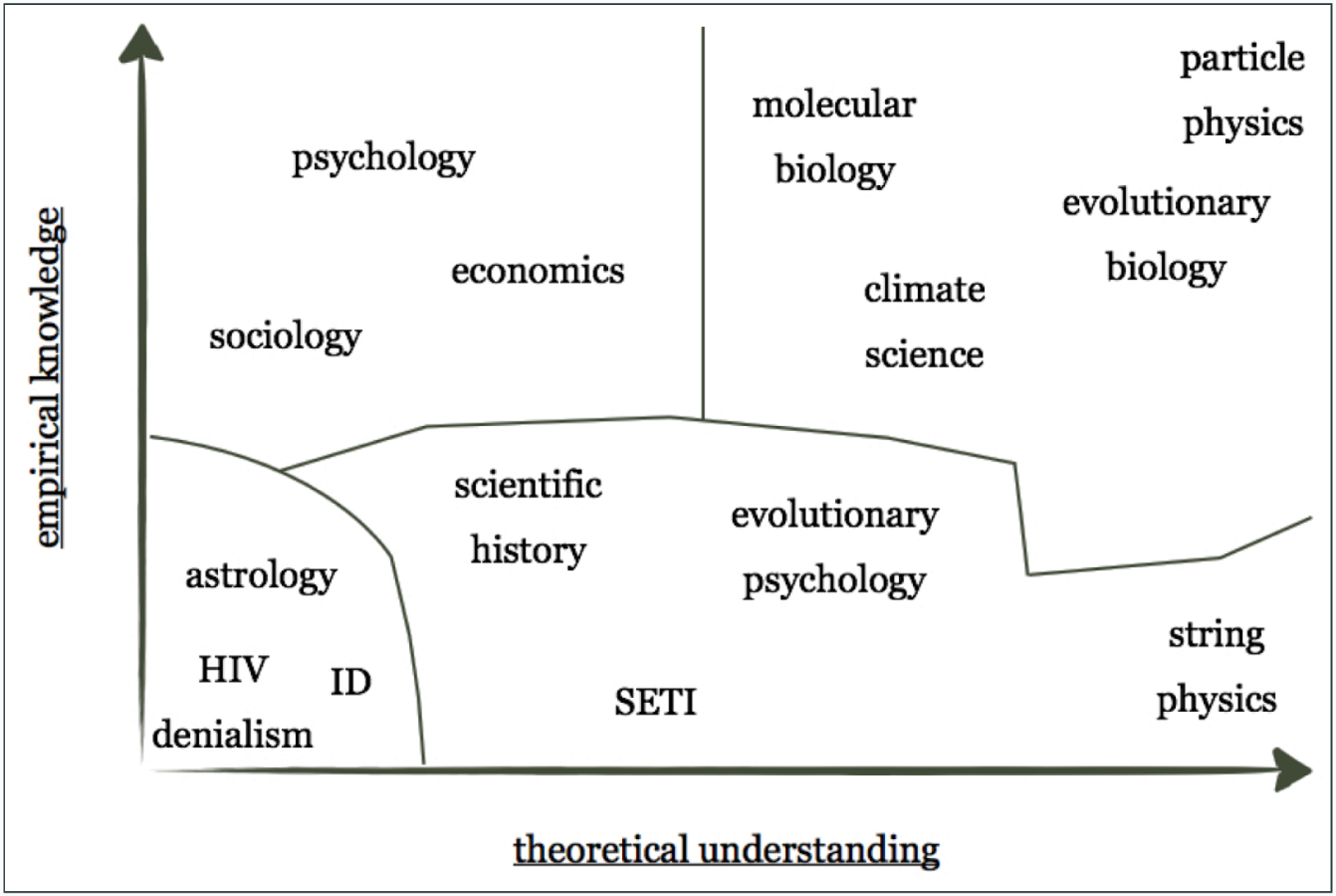
→ Employs empirical reasoning to transform data into knowledge

→ Relies on notions and tools from scientific methodology

The “scientificity” cursor

- Today – less binary view:
  - → The “scientificity” of a field is seen as a spectrum
  - → Some fields are more scientific, others rather pseudo-scientific
- Certain criteria, although imperfect, remain strong indicators
  - Scientific method
  - Testability, reproductibility
  - Calling into question vs. dogmatic beliefs

Theoretical vs. empirical understanding



Source: Pigliucci & Boudry (2013), *apud* Le Chat Sceptique

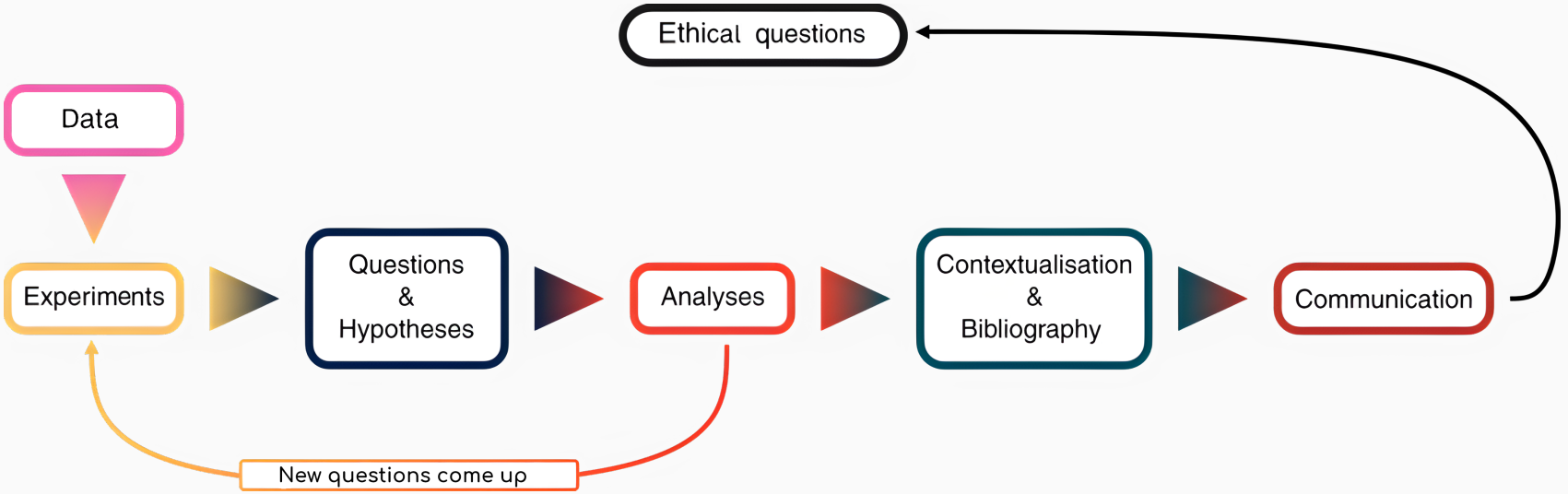
Epistemo. . . what?

Congratulations! We just practised a bit of epistemology!

It is the science of knowledge

Outline

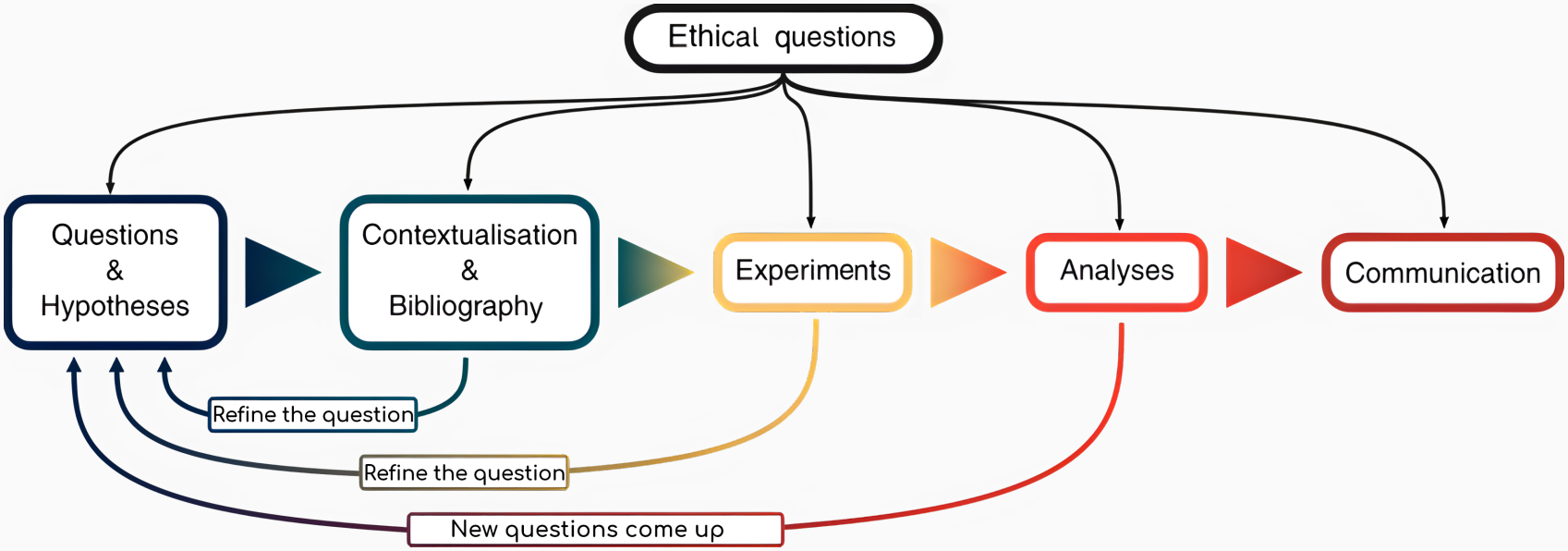
The method: in practice



Wooclap

Wooclap time!

The method: ideally



Beginning of the project, generally chaotic...

- At the start of a project, we may feel lost:
  - → a vague idea of the subject
  - → ...but without knowing how to go about it
  - → ...and fearing that the work might start off wrong
- The original chaos: a stressful but unavoidable situation
- Scientific methodology is:
  - → A journey towards a better understanding
  - → Something that by definition is "being searched for"
  - → Thus, it includes hesitations, drifts, and uncertainties

Traps of the original chaos 1/4

- Overlook the hypotheses
  - **The "hype"**
    - ⇒ Headlong rush driven by the belief that using sophisticated techniques increases the intellectual value of a work
    - ⇒ Being trendy: testing the latest technological advances
    - ⇒ *"I'll apply all available machine learning models in this library, and then we'll see. . . "*

Traps of the original chaos 2/4

- Overlook the hypotheses
  - **The data tsunami**
    - ⇒ Collecting the data before formulating the hypotheses
    - ⇒ A societal issue: faster, better, stronger, data everywhere!
    - ⇒ *"Here's my database, do some machine learning on it."*

Traps of the original chaos 3/4

- Overlook the hypotheses
  - **Scientific engineering**
    - ⇒ Develop tools/resources that are useful
    - ⇒ Address a practical issue, not a knowledge gap
    - ⇒ *"I spent so much time building the perfect software, it must answer some question!"*

Traps of the original chaos 4/4

- Too many ideas, not enough me's
  - $\Rightarrow$ Excessive ambition $\rightarrow$ absolute confusion
  - $\Rightarrow$ *"I'm going to create a perfect machine translation system that works for all languages in the world."*

Research question: why does it matter?

The research question helps to overcome the original chaos

- Clarifying your intentions
- Formulate the project in the form of a question
- This question attempts to express what we want to know or understand better

$\Longrightarrow$ First common thread of a scientific project

Problematise

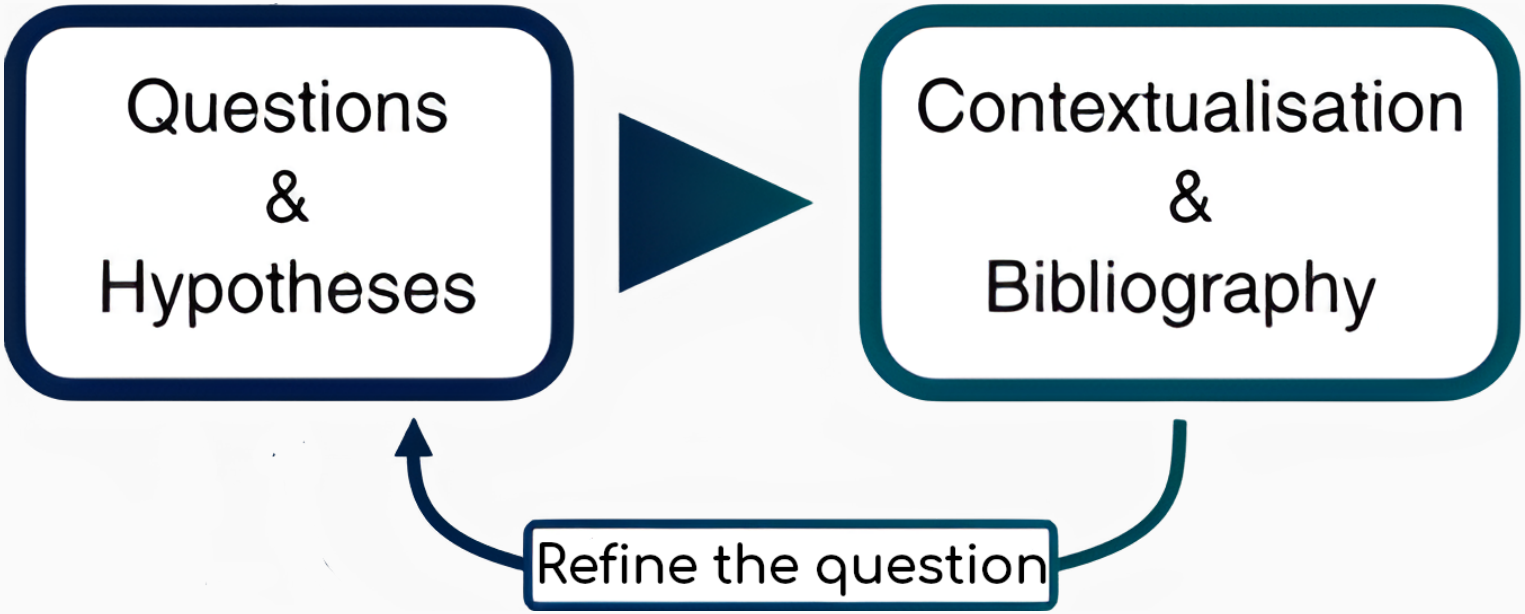
- Identify a problem/issue:
  - Gap: current unsatisfactory situation vs. target situation
  - Work undertaken must allow reducing this gap by answering a research question
- Problematise (raise questions):
  - Proceed from a general topic to a research question
  - Based on mix: bibliography + personal experience/intuition

    $\rightarrow$ Iterative process

Analyse: slicing up the mountain

- Transforming a huge, vague problem...
- ...into a series of simpler, more precise problems

Chicken and egg situation



Problematise: how, in which context?

Issue

Marine heatwaves are becoming increasingly frequent. Their impact on the diet of marine creatures, particularly common octopus, remains unknown.

Research question

What impact do marine heatwaves have on the quantity of food ingested by common octopus (Octopus vulgaris)?

Bibliography

Identifies a knowledge gap related to the issue

Existing articles measure the impact of heatwaves on small samples of octopus, over short periods of time, and only on their weight, not on their diet.

Research question

Research question

What impact do marine heatwaves have on the quantity of food ingested by common octopus (Octopus vulgaris)?

Goals

- Measure the daily food intake of numerous octopuses over several years
- ...
- Compare these values between days with/without heatwave
- Compare these values between summer and winter

Hypotheses

- The amount of food ingested by octopuses during periods of marine heatwave decreases
- The distribution of food eaten in each season (summer vs. winter) changes in heatwave years

Research question

Main research question

What impact do marine heatwaves have on the quantity of food ingested by common octopus (Octopus vulgaris)?

Goals

- Measure the daily food intake of numerous octopuses over several years
- ...
- Compare these values between days with/without heatwave
- Compare these values between summer and winter

Hypotheses

- The amount of food ingested by octopuses during periods of marine heatwave decreases
  True/false ?
  Secondary question
- The distribution of food eaten in each season (summer vs. winter) changes in heatwave years
  True/false ?
  Secondary question

Secondary question/sub-question: shorter and less complex

Research question

Research question

Does the quality of machine translation (measured by the BLEU score on a standard English-French dataset) increase when more data is available to learn a translation system based on Transformers?

Goals

- Train a French-English MT system on 10k sentences
- Train a French-English MT system on 20k sentences
- ...
- Compare the BLUE scores of the systems, observe the trend

Hypotheses

- The BLUE score will be higher when more training data is used
- The increase in the BLUE score is not linearly proportional to the amount of data added

Research question

Main research question

Does the quality of machine translation (measured by the BLEU score on a standard English-French dataset) increase when more data is available to learn a translation system based on Transformers?

Goals

- Train a French-English MT system on 10k sentences
- Train a French-English MT system on 20k sentences
- ...
- Compare the BLUE scores of the systems, observe the trend

Hypotheses

- The BLUE score will be higher when more training data is used
  True/false ?
  Secondary **question**
- The increase in the BLUE score is not linearly proportional to the amount of data added
  True/false ?
  Secondary **question**

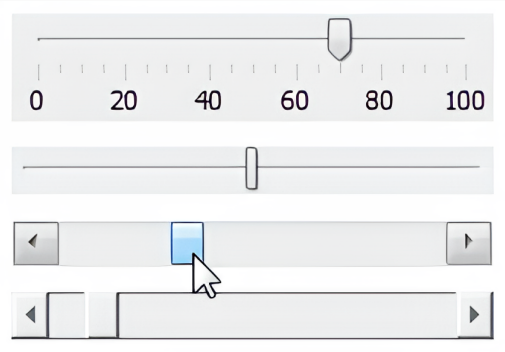
Secondary question/sub-question: shorter and less complex

Answers to the research question

- The secondary questions almost completely define the experiment(s) to be carried out
- Most often: no definitive answer to the main question
- Answers are given to the secondary questions, and little by little we find clues to the answer to the main question

Belief slider

- Answers to secondary questions adjust the belief slider of the main question
- The slider on the main question may never reach 0 or 1



What makes a GOOD research question?

Moving from a research idea to a research question

Idea	Research question
Creative	Systematic
Impulsive	Well thought through
Thrilling	Measured
Vague	Formalised
Not verified	Verifiable
Not validated	Validated
Unlimited	Time and budget constraints

Source: Martinez and Berkhout (2009)

What makes a GOOD research question?

What makes a GOOD research question?

- Relevant

What makes a GOOD research question?

- Relevant
- Focused and clear

What makes a GOOD research question?

- Relevant
- Focused and clear
- Feasible

$\longrightarrow$ Leading to a realistic protocol

Relevance

- Has my study not already been carried out?
  - $\rightarrow$ If yes, will it provide missing results?
  - $\rightarrow$ If yes, will it confirm uncertain information?
- Will my study lead to progress in my field of expertise?
  - $\rightarrow$ Is it likely to have an impact on society?
- Can I find some personal interest and motivation in the study?

$\Rightarrow$ Question the impact of the research

Relevance

- Have I clearly defined the scope of my study?
- Do I have a good knowledge of the field of my study and the related literature?
- What are the important research questions in this field?
- What areas of the field of study deserve further exploration?
- Will my study fill an information gap? Will it provide a better understanding of the field of study?
- Have there not already been many studies carried out in this field?
- Is this the right time to answer this question? Is it not too burning a question or, on the contrary, one that will quickly become obsolete?

Wooclap

Wooclap time!

Focus and clarity

A clear research question is:

- **Precise:** its meaning is not misleading

Counter-examples

What are the impacts of heat on marine creatures?

⇒ What type of heat? Summer heat, or heatwaves? Water or air heat? What sea creatures?

What are the impacts of recent AI improvements on NLP?

⇒ What improvements? In which AI techniques? Which NLP areas or applications?

Check the understanding of several people ⇒ identical or divergent?

Wooclap

Wooclap time!

Focus and clarity

A clear research question is:

- **Concise:** remember the beginning when you get to the end
- **Unambiguous:** only one possible interpretation
  - The terms used are defined beforehand
  - The wording is as comprehensible as possible

Counter-examples

To what extent do episodes of intensely excessive heat, whether due to climate change-related phenomena or simply due to fluctuations linked to the substantial water temperature differential between the warm and cold seasons, have an impact on the daily and seasonal consumption of food by individuals of Octopus Vulgaris?

To what extent can we assume that the increase in the amount of corpus-based textual data aligned at sentence level contributes to performance changes, either improvements or losses, in machine translation systems based on modern neural models predicting the target sentence using an auto-regressive paradigm?

The question is too long and confusing

Wooclap

Wooclap time!

Focus and clarity

A clear research question is:

- **Focused:** often, the initial question is too broad

Counter-examples

Why do octopus diets change so much in summer?

Why does the quality of MT systems vary so much across languages?

- Avoid starting a main research question with "why"

  → (Too) many possible answers

- Avoid subjective terms

  → E.g. *so much*, *in general*, *frequently*, *a lot*, *few*, ...

Focus and clarity

Often, the initial question is too broad

Detailed example:

Does having more data available improve the performance of machine translation systems?

Focus and clarity

Often, the initial question is too broad

Detailed example:

Does having more data available improve the performance of machine translation systems?

$\Rightarrow$ What does "improve" mean here? What exactly is the aim?

Different **metrics** focus on different aspects of the results: precision, recall, accuracy, AUC, F1, ...

Focus and clarity

Often, the initial question is too broad

Detailed example:

Does having more data available improve the performance of machine translation systems?

$\Rightarrow$ Which type(s) of machine translation system(s)? Transformer? Statistical? Rule-based? ...

Different **systems** may be more or less dependent on the amount of data available

Focus and clarity

More targeted/focused examples

What impact do marine heatwaves have on the quantity of food consumed by common octopus (Octopus vulgaris)?

Does the quality of a Transformer-based MT system, measured by BLEU score on a standard English-French dataset, increase with more training data?

Focus and clarity

Question too broad / vague:

$\Rightarrow$ Must be narrowed down

$\Rightarrow$ A common and frustrating process for young researchers

1. Bibliography $\rightarrow$ identify the questions waiting to be answered
2. Formulate several candidate research questions
3. Choose the most relevant and focused

Focus and clarity

Question too broad / vague:

$\Rightarrow$ Must be narrowed down

$\Rightarrow$ A common and frustrating process for young researchers

1. Bibliography $\rightarrow$ identify the questions waiting to be answered
2. Formulate several candidate research questions
3. Choose the most relevant and focused

Beware of excessively focused and restrictive questions

$\rightarrow$ Limited impact

Feasibility

Do I have the means needed to answer my research question?

- Data collection (availability of sources, equipment. . .)
- Technical expertise
- Time
- Funding
- Team and team management

Time and material constraints are often underestimated!

Feasibility

Can I find funding sources?

Are there organisations interested in my work that could fund it?

Will my study have an impact in its field and on society?

Can I convince funding bodies that I am capable of carrying out this research?

- Expertise
- Preliminary results
- History of succesful projects

“Our” framework - QHJE

Our proposal for structuring the work

- Question – multi-level, with precise sub-questions
- Hypotheses – associated to each sub-question
  - If I get this result, my conclusion will be X, otherwise Y
- Justification – my research questions are
  - relevant,
  - potentially useful,
  - not yet covered in the literature
- Experiment design – associated with each sub-question
  - Datasets
  - Experimental conditions
  - Evaluation metrics
  - ...

Example: octopus and heatwave

Research question: *What impact do marine heatwaves have on the quantity of food consumed by common octopus?*

Question Q1: Does the amount of food ingested by octopuses during periods of marine heatwave decrease?

- **Hypothesis** H1: Yes
  - H1 is confirmed if, on average, octopuses eat significantly less during heatwave periods
- **Justification** J1: Humans eat less when the weather is hot
- **Experiment** E1: Compare average food quantity during heatwave vs. without heatwave (1 week before)

Example: octopus and heatwave

Research question: *What impact do marine heatwaves have on the quantity of food consumed by common octopus?*

Question Q2: Does the distribution of food eaten in each season (summer vs. winter) change in heatwave years?

- **Hypothesis** H2: Yes
  - H2 is confirmed if the average quantity of food consumed by octopuses in winter is significantly greater in heatwave years
- **Justification** J2: Octopuses compensate for periods of "heatwave starvation" by eating more in the winter
- **Experiment** E2: Compare average food quantity in summer vs. winter in years with/without marine heatwaves

Granularity of a research question

- The level of specificity (granularity) of a research question depends on the context of the work
  - $\rightarrow$ **Thesis:** general research question, with very specific sub-questions
  - $\rightarrow$ **Article:** specific research question for a more focused contribution

Warning!

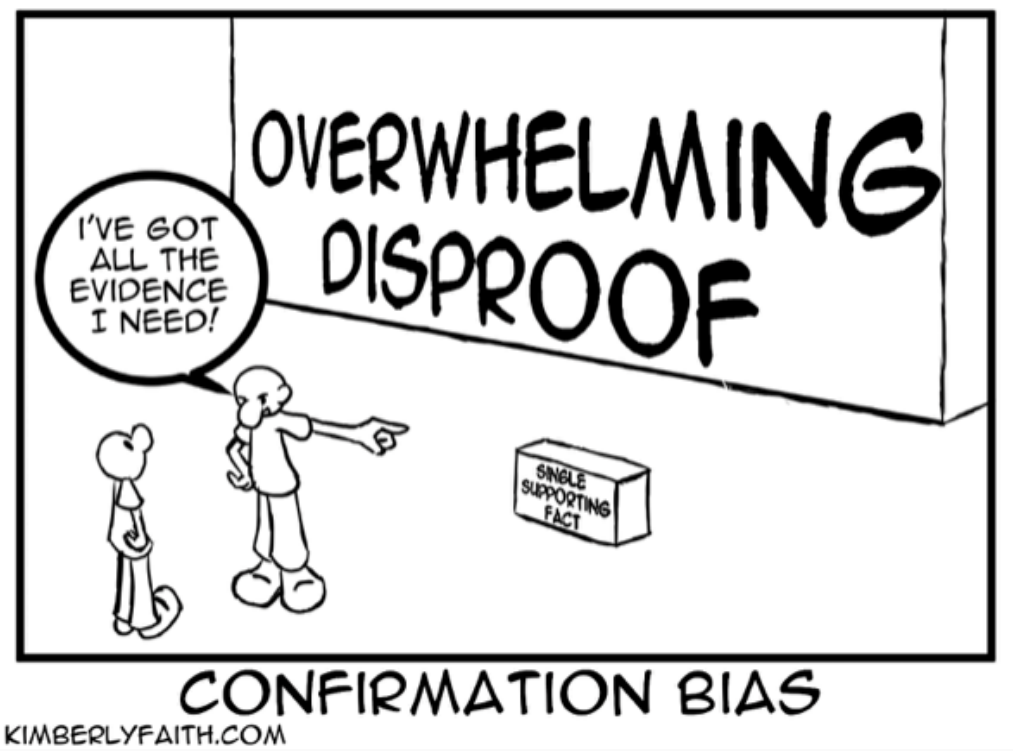
What is a cognitive bias?

- Systematic pattern of deviation from norm or rationality
  - → Biased stimuli perception → "subjective reality"
  - → Behaviour that can distort the results of a research
- May lead to:
  - → Distorted perception
  - → Inaccurate reasoning
  - → Illogical interpretation
  - → Irrationality

Source: `https://en.wikipedia.org/wiki/Cognitive_bias`

Confirmation bias

- Tendency to favour interpretations that confirm initial beliefs
  - $\rightarrow$ Only search for what we want to find
- May lead to cognitive dissonance, well studied in psychology
  - *For tagging, accuracy increases when using WALS. So it works!*
  - *For parsing, accuracy decreases, but for this particularly hard language it increases. So it works!*



Outline

I've got my research question! . . . What's next?

It is time to take a closer look at the work of our colleagues.

It is time to make:

- A bibliographic search!
- A literature review!

Well, it's the same thing... No?

What is the difference?

- Goal of a bibliographic search
  - → **Acquire and expand** our knowledge of a specific subject
- Goal of a literature review (survey)
  - → **Summarise the state of the art** of what is known in the field

What is the difference?

⚠

- A bibliographic search is NOT a literature review
- But a literature review implies doing a bibliographic search

When should we do either?

- Bibliographic search :
  - $\longrightarrow$ Before any research project
  - $\longrightarrow$ During the refinement of the research question
- Literature review (survey):
  - $\longrightarrow$ The last survey is out of date OR
  - $\longrightarrow$ Such a review has never been done AND
  - $\longrightarrow$ I have time to sit down and review the state of the art

How should we do either?

- Bibliographic search :

  $\longrightarrow$ No strict method

  $\longrightarrow$ Idea of starting point: one article leads to read another article, like a dictionary where a first definition leads to a second, a third, and so on.

- Literature review (survey):

  $\longrightarrow$ **Narrative** review: no strict method

  $\longrightarrow$ **Systematic** review: follow existing methodology (e.g. PRISMA)

  $\longrightarrow$ Search in priority for already existing surveys, meta-analyses

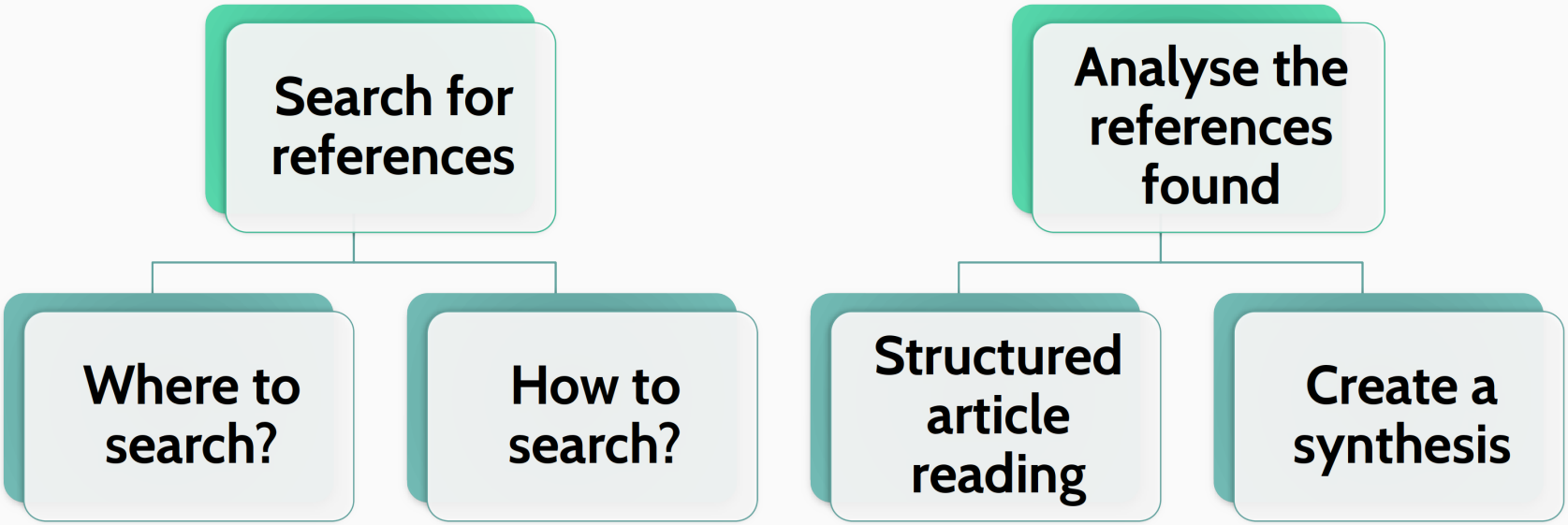
  $\longrightarrow$ Define (a priori or as we go along) the scope of the review

  $\longrightarrow$ A survey will generally be published

Wooclap

Wooclap time!

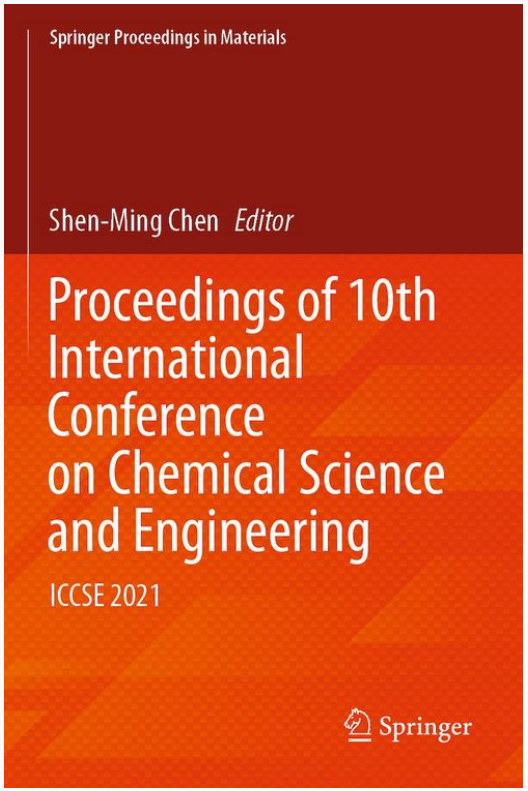
The process



Where to search?

- Journals and conference proceedings

  $\longrightarrow$ University libraries often have subscriptions to non-free journals



Where to search?

→ Caution: no peer reviewing

Where to search?

Repositories

- National: in France, HAL
- International: generally managed by universities
- Specialised: DBLP (informatics), ACL Anthology (NLP)...

Where to search?

- Other sources:
  - Web pages of laboratories, teams and researchers in the field
    $\rightarrow$ Sometimes you can find "preprints" on personal websites
  - Twitter (X)  , LinkedIn  , social networks (some fields),...
    

How to search?

- Google Scholar, Social Science Research Network (SSRN), Semantic Scholar, Scinapse, ...
  - Choosing the right keywords
  - Identify the most important articles
    - $\rightarrow$ Number of citations and downloads

How to search?

- Google Scholar, Social Science Research Network (SSRN), Semantic Scholar, Scinapse, ...
  - Choosing the right keywords
  - Identify the most important articles
    - $\rightarrow$ Number of citations and downloads
- Newsletters from journals and publishers
  - IEEE, Springer, Elsevier, ...

How to search?

- Google Scholar, Social Science Research Network (SSRN), Semantic Scholar, Scinapse, ...
  - Choosing the right keywords
  - Identify the most important articles
    - $\rightarrow$ Number of citations and downloads
- Newsletters from journals and publishers
  - IEEE, Springer, Elsevier, ...
- E-mail alerts on keywords: Google Scholar, ArXiV, ...

How to search?

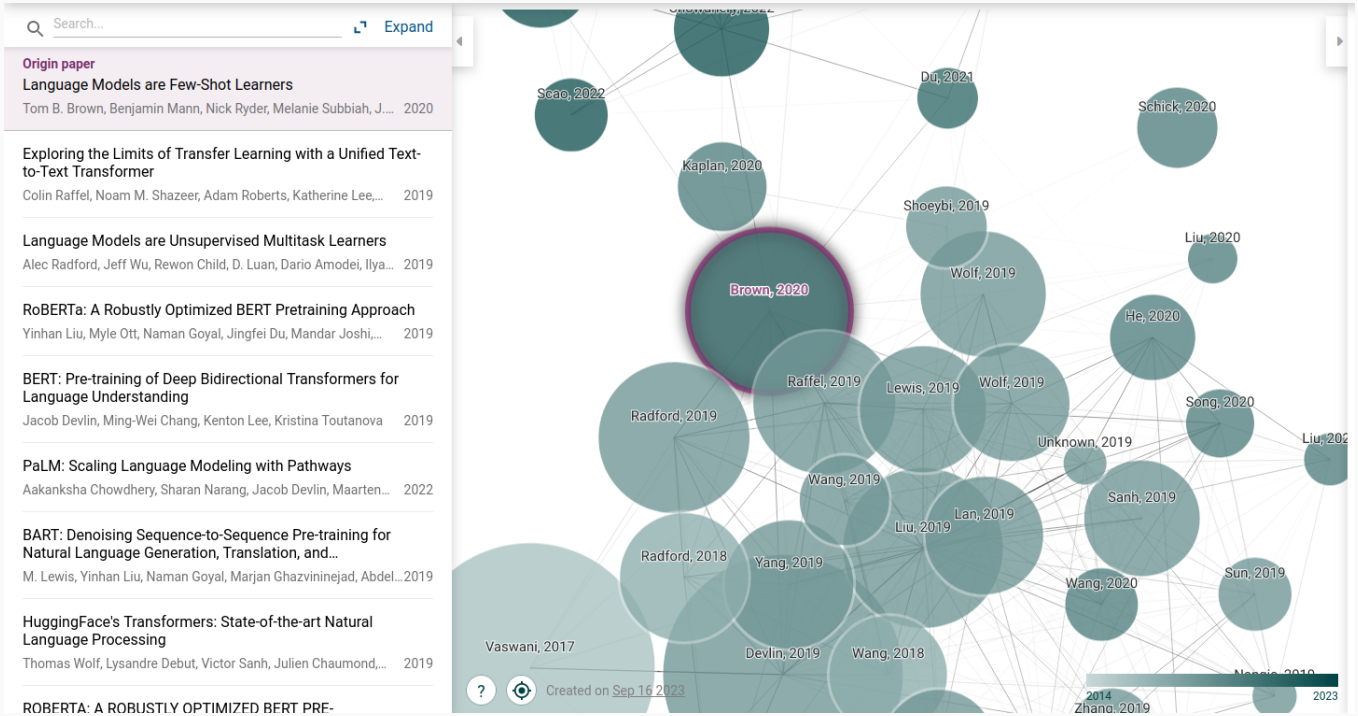
- Google Scholar, Social Science Research Network (SSRN), Semantic Scholar, Scinapse, ...
  - Choosing the right keywords
  - Identify the most important articles
    - → Number of citations and downloads
- Newsletters from journals and publishers
  - IEEE, Springer, Elsevier, ...
- E-mail alerts on keywords: Google Scholar, ArXiV, ...
- Thematic and scientific associations mailing lists
  - → National projects, EU projects (e.g. COST) ...
  - → Conference proceedings and journal issues announcements

How to search?

- Google Scholar, Social Science Research Network (SSRN), Semantic Scholar, Scinapse, ...
  - Choosing the right keywords
  - Identify the most important articles
    - → Number of citations and downloads
- Newsletters from journals and publishers
  - IEEE, Springer, Elsevier, ...
- E-mail alerts on keywords: Google Scholar, ArXiV, ...
- Thematic and scientific associations mailing lists
  - → National projects, EU projects (e.g. COST) ...
  - → Conference proceedings and journal issues announcements
- Following links
  - Found in literature reviews of related work

How to search?

- Research Rabbit, Connected Papers...
  - → Article visualisation tools
  - → Research Rabbit: free, but requires signing up[1]
  - → Connected Papers: 2 free searches, then paywall



[1] This was verified in June 2024, but business models change quickly.

Finding references

Bibliographic references must be:

- Useful
- Recent
- Original

Finding references

How to find useful references?

- Beware of the dates!
  - Full chronological search: all
  - Search for the latest advances: $<$ ∼5 to 10 years
- Keep in mind that we will not be able to read everything!
  - Prioritise by relevance and impact
  - Order of magnitude: ∼50-100 articles for a PhD
    - $\rightarrow$ It is not necessary to read each article in its entirety
    - $\rightarrow$ Nor with a perfect level of understanding!
  - Stop when new articles seem predictable
    - $\rightarrow$ sufficient knowledge of the field

Finding references

How to find recent references?

- What counts as recent depends on the field
  - → Cognitive psychology: ∼10 last years
  - → Machine learning: ∼3 last months

Example: ACL reviewing policies

If you are aware of relevant publicly available research that has not been cited [...], you should bring it to the attention of the authors [...] However, if the work appears only in a preprint, especially one that is recent and/or not widely cited, you should [...] not penalize them [...] it is ***not reasonable*** *to expect a* ***time-consuming empirical comparison*** *with work that has appeared* ***less than 3 months*** *before the submission deadline.*

Wooclap

Wooclap time!

Finding references

How to find original references?

- Avoid reading always the same authors and labs
  - Cover different methods and currents of thought
- Remain open-minded
  - Inspiring ideas can come from where you least expect them
  - Going to conferences, reading groups, seminars
  - Multidisciplinarity: draw inspiration from other fields
- Little by little, build up a broad knowledge basis

Analysing the references found

For each reference, follow 5 steps below:

The structure of a scientific article

- **Abstract**

  → Presents the main information, very short (1 paragraph)

1. **Introduction**

   → Context, problem, research question and hypothesis

2. **Related work**

   → What do we know up to now? What do we do that others don't?

3. **Methodology**

   → Description of proposed method, model, experiment, algorithm...

   → Experimental protocol and data

4. **Experiments results**

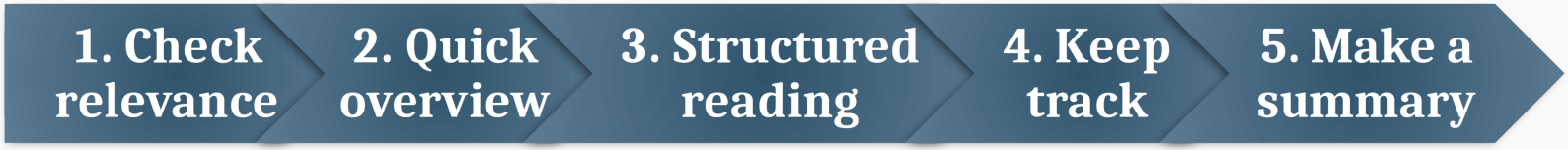
   → Obtained results (metrics, tables, graphs)

   → Discussion and comparison with previous work

5. **Conclusions**

   → Summarise: what did we learn with this work? What's next?

1. Check relevance



- Read the sections
  - Abstract
  - Introduction
  - Conclusions
- Overview the other sections
- Is the theme relevant to my study?
- Have the goals been achieved?

2. Quick overview

- Quick reading of the article without delving into difficult points

Focus on:

- Summary, introduction
  → To understand the subject
- Figures (images), results tables (numbers) and conclusion
  → Preliminary idea of the achieved outcomes
- Desired goal :
  → Overall comprehension of the article
  → Can we read it in detail right away, or is some preparation required?

3. Structured reading

- Detailed reading, complements the global overview
- Search for additional resources if necessary:
  - Articles cited
  - Methods cited
  - Previous work by authors
  - ...
- Identify:
  - Aim / topic and research question(s)
  - Methods used
  - Comparative results: strengths and weaknesses
  - Take notes

4. Keep track of the work

1. Check relevance
2. Quick overview
3. Structured reading
4. Keep track
5. Make a summary

- List of references
  - Bibliography management tools: Zotero, Mendeley, JabRef...
- Reading notes
  - For each article / resource
  - Choice of format: text file, spreadsheet, paper/ pen...
  - ...
- Brief reports

4. Keep track of the work

1. Check relevance
2. Quick overview
3. Structured reading
4. Keep track
5. Make a summary

- Citations with BiBTeX - Overleaf
  - → Automatic formatting and sorting
  - → Available on most platforms (export BiBTeX)
- "Cleaning" tools: `bibclean`
- Autocomplete on overleaf and similar LaTeX editors

BiBTeX: example

```
@article{smith2020marine,
  title={Marine Heatwaves: An Emerging Global Threat},
  author={Smith, Karen and Jones, Robert and Williams, Sarah},
  journal={Nature Climate Change},
  volume={10},
  number={1},
  pages={12-17},
  year={2020},
  publisher={Nature Publishing Group}
}
```

5. Make a summary

Purpose of the summary:

- ~~Pile up articles, list them independently of each other~~
- Identify different approaches to study the question
- Group items according to common characteristics
  - Theories
  - Methodologies used
  - Models or algorithms used
  - Data, datasets
  - Currents of thought
  - Conclusions obtained
  - ...

It is up to you to identify the relevant characteristics!

5. Make a summary

1. Check relevance
2. Quick overview
3. Structured reading
4. Keep track
5. Make a summary

- What to say in a summary?
- The different categories identified
- The articles which fit into these categories
  - Why these articles belong (or not) to the categories
  - What variants they introduce
- Critical analysis
  - Points of comparison between the different categories
  - Show the limits of previous work: aspects missing or unsatisfactory in relation to our research question

5. Make a summary

Method:

1. Read articles and take notes
2. Choose the categorization(s) to use
   - Possibility of organising them by logical progression (e.g. deep learning, dependency parsing, multilingual methods...)
3. Identify the main sub-categories/variants (e.g. curriculum learning, contrastive learning, instruction fine-tuning...)
4. Present the articles in each category and sub-category
   - With citation '[x]' or '(Doe 2016)'

5. Make a summary

1. Check relevance
2. Quick overview
3. Structured reading
4. Keep track
5. Make a summary

To write a good summary:

- Not a catalogue of articles with no links between them
  - → In 1-2 sentences: summarise what is important in the article
  - → Focus: aspects relevant to my work (similar/different)
  - → Use writing aids: Thesaurus, Ref-n-Write, ...

5. Make a summary

To write a good summary:

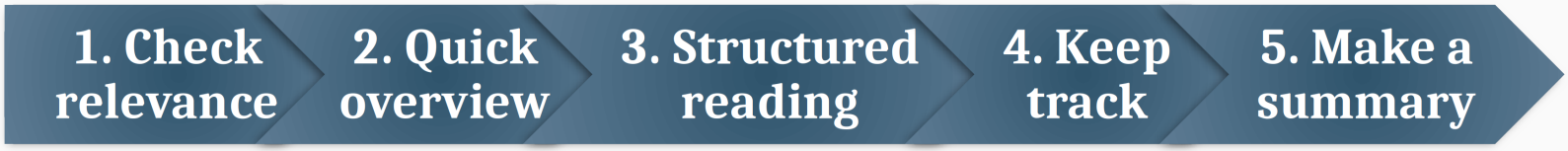
- Not a catalogue of articles with no links between them
  - → In 1-2 sentences: summarise what is important in the article
  - → Focus: aspects relevant to my work (similar/different)
  - → Use writing aids: Thesaurus, Ref-n-Write, ...
- Stay focused on the main topic
- Provide the essential information
  - → Not too many details for each article
  - → This is not a course and you are talking to experts

5. Make a summary

To write a good summary:

- Know who you are talking to
  - → In general: colleagues in the same field

5. Make a summary



To write a good summary:

- Know who you are talking to
  - → In general: colleagues in the same field
- Always indicate your sources
- Cite "respectable" sources
  - → Scientific publications, official reports...
  - → Beware of too recent, unreviewed articles (arXiv)

5. Make a summary: counter-example

CITATION1 showed that XXX. CITATION2 used Y1 and showed that this technique is the most effective.

Y2 was used by CITATION3 who highlighted XXX as an advantage. CITATION4 preferred to use Y1, as did CITATION2 and CITATION5.

5. Make a summary: counter-example

CITATION1 showed that XXX. CITATION2 used Y1 and showed that this technique is the most effective. → Most effective in what context? Compared to what?

Y2 was used by CITATION3 who highlighted XXX as an advantage. CITATION4 preferred to use Y1, as did CITATION2 and CITATION5. → Why?

5. Make a summary: exemple

This work is at the intersection of three trends in the literature on YY. The first is Y1, which consists of XXX. The advantage of this method is that XXX. The second is Y2, which aims to XXX. This second method has the advantage of XXX. The third and final trend is the use of Y3, which offers a wide range of information on XXX.

Y1 is an effective solution when XXX. CITATION describe two types of Y1. The first is based on XXX. For example, CITATION proposes XXX. These methods sometimes produce unexpected results: XXX. The authors' hypothesis is that XXX.

5. Make a summary: exemple

This work is at the intersection of three trends in the literature on octopus feeding. The first is stuffing, which consists of eating as much as possible in a short time. The advantage of this method is that the octopus creates fat reserves. The second is storage, which aims to create food reserves in its den. This second method has the advantage of spreading out the food over time. The third and final trend is the use of tools, which offers a wide range of information on these cephalopod's hunting techniques.

Stuffing is an effective solution when the octopus has hunted a large prey. CITATION describe two types of stuffing. The first is based on the opportunism of the octopus. For example, CITATION proposes to lay open oysters in front of an octopus's den. These methods sometimes produce unexpected results: although easily accessible food was available, more than half the octopuses chose to retreat to their den. The authors' hypothesis is that these octopuses are trying to avoid potential danger.

5. Make a summary: exemple

This work is at the intersection of three trends in the literature on syntactic parsing in multilingual dependencies. The first is transfer parsing, which consists of learning a parser on one language (or a set of languages) and then testing it on another. The advantage of this method is that it can be applied to languages with few or no resources, since all that is needed is training data for the source language. The second is delexicalised parsing, which aims to ignore the lexicon. This second method has the advantage of neutralizing the biases of textual genre or domain, which are strongly marked in the vocabulary of the corpora. The third and final trend is the use of typological resources, which offers a wide range of information on languages.

Transfer parsing is an effective solution when dealing with languages with few resources. CITATION describe two types of transfer parsing. The first is based on parallel corpora, when one of the two languages has no, or not enough, training data. For example, CITATION proposes a parser for Irish first trained on another language, then applied to Irish. These methods sometimes produce unexpected results: although they do not belong to the same language family, Indonesian gives the best results in this approach. The authors' hypothesis is that distant dependencies are better represented in Indonesian than in the other languages tested.

“Our” framework: bibliography

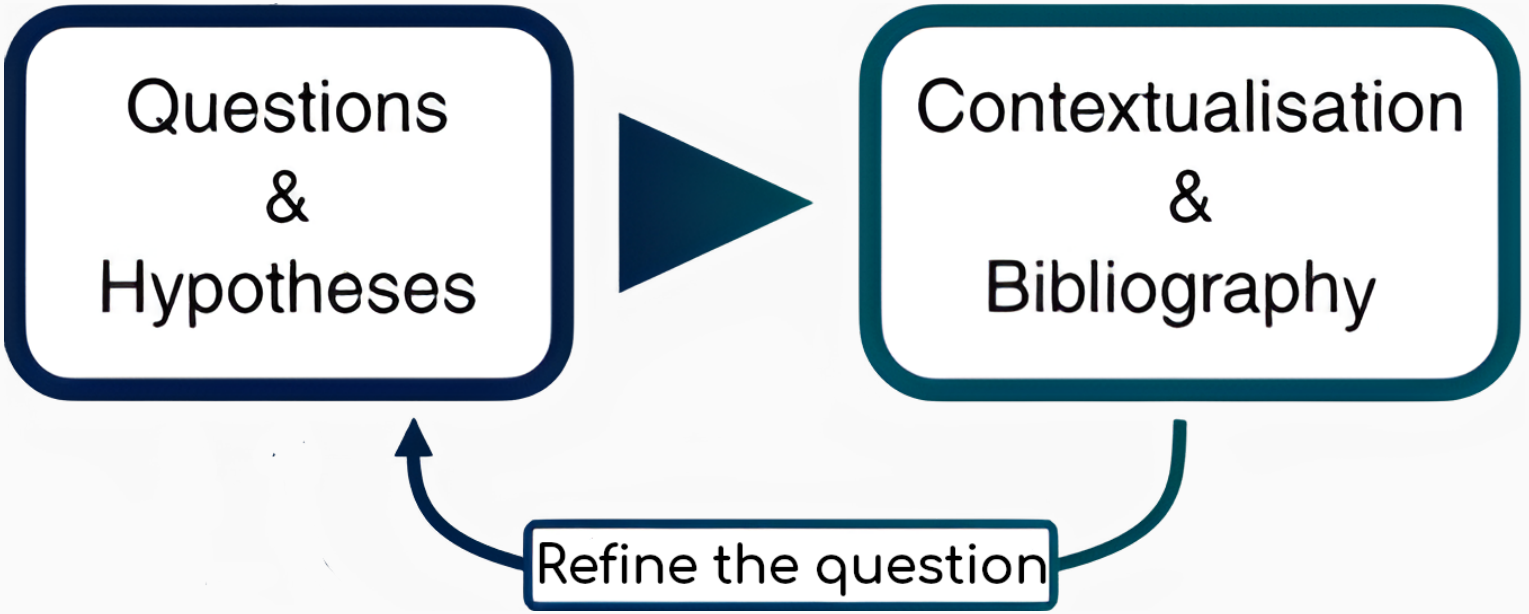
A proposal for structuring the work:

- **Title, Author**
- **Abstract**
- **Contributions**
- **Similarities** with my work
- **Differences** with my work
- **Comments**

“Our” framework: bibliography

Paper	Abstract	Contributions	Similarities	Differences	Comments
Wang and Eisner (2018)	This article looks at the use of surface statistics for parsing an unknown language. . .	- Features learned from POS-annotated corpora help parsing - Using “synthetic” languages during training improves results - . . .	- Delexicalised - Zero-shot - . . .	- Unsupervised - They do not use any parallel data - . . .	- Their system depends on gold POS tags - Interesting criticism of WALS: - *The unknown language might not be in WALS* - . . .
de Lhoneux et al. (2018)	The aim of this article is to test different strategies for sharing the parameters of a dependency parser. They test . . .	- Some parameters are useful for information sharing, others not - The MLP used as a classifier - . . .	- Multilingual - Zero-shot - . . .	- Lexicon - Semi-supervised - . . .	- They test 27 different configurations - Their tests are conducted by training 5 bilingual models - . . .
. . .	. . .	. . .	. . .	. . .	. . .

Reminder: iterative process



Bibliography: in short

- First step in any research project
- Iterative process - definition of the research question
- Goals
  - Building an scientific argumentation
  - Justify the research question
    - $\longrightarrow$ Relevant - filling a gap
    - $\longrightarrow$ Feasible - building on what already exists
    - $\longrightarrow$ Interesting - potential impact in the field

Thanks!

That's all for today

Carlos Ramisch
first.last@lis-lab.fr

M2 IAAA - based on the course *Zen Research*
By Carlos Ramisch and Manon Scholivet

Sources i

- Adeline Paiement's course *Initiation à la recherche*
- Damien Driot's course *Débuter un travail de recherche*
- Pigliucci & Boudry (2013) *Philosophy of Pseudoscience: Reconsidering the Demarcation Problem*
- Sartenaer (2020). *Différencier science et pseudoscience : pas si simple*
- Youtube channels: *Le Chat Sceptique*, *Hygiène Mentale*, *Monsieur Phi*
- Feedback from participants of previous course editions

Sources ii

- Slides illustrated with the help of: Jérémy Trione, Google images, Midjourney, `imgupscaler.com`, Canva
- Slides written with the help of: ChatGPT, Google Bard, DeepL, Linguee, Overleaf
- Funding: French ANR, through SELEXINI project (ANR-21-CE23-0033-01)

Backup slides

Literature review (survey): example

Question: *Methodological choices made in the experiments?*

Scope

- Articles >2016 using datasets X or Y, available on Z

Analysis table

- 2. Data
  - → 2.1 Source (news, wikipedia, social media...)
  - → 2.2 Languages
- 3. Data processing
  - → 3.1 Pre-processing
  - → 3.2 Post-processing
- 4. Evaluation
  - → 4.1 Metrics (exact match, fuzzy match...)
  - → 4.2 Significance tests

Source: Based on `https://aclanthology.org/2023.mwe-1.15/`

Literature review (survey): example

	2 Languages	3 Split of the corpora	3.4 Category of	4.1 Preprocessi
PARSEME 1.0				
The PARSEME Shared Task on Autom	18: BG, CS, DE, EL,	train/test, no dev		N/A
Parsing and MWE Detection: Fips at th	8: FR, EN, DE, IT, E	Not mentioned	VID, LVC, VPC,	Transformation b
The ATILF-LLF System for Parseme Sh	18: BG, CS, DE, EL,	PARSEME data	PARSEME categ	Not mentioned
Detection of Verbal Multi-Word Express	15: CS, DE, EL, ES,	PARSEME data	VPC, LVC, VID,	Not mentioned
USzeged: Identifying Verbal Multiword	9: DE, EL, ES, FR, H	PARSEME 1.0 (no dev	PARSEME 1.0 c	Remove long se
A data-driven approach to verbal multiv	12: RO, FR, CS, DE,	PARSEME 1.0 - cross	PARSEME 1.0 c	Not mentioned
Neural Networks for Multi-Word Expres	15: BG, CS, DE, EL,	80% train, 10% dev, 1(	PARSEME 1.0	Not mentioned
PARSEME 1.1				
Edition 1.1 of the PARSEME Shared Ta	19: BG, DE, EL, EN,	3 languages had no de	LVC, VID, IRV, V	N/A
CRF-Seq and CRF-DepTree at PARSE	19: BG, DE, EL, EN,	PARSEME 1.1 data	PARSEME 1.1	Converting to XM
Deep-BGT at PARSEME Shared Task 2	10: BG, DE, ES, FR,	PARSEME 1.1 data	All PARSEME 1.	Merging labels, r

Literature review: in short

- Synthesis of the state of the art in the field
- Requires a structured reading of a large number of articles
  - → Ability to synthesise
  - → Comparison, putting into perspective
  - → Analysis and structuring of content
  - → Identifying the challenges and open problems of the field
- Valorised / published in the form of a survey or meta-analysis